# MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine Learning algorithms are categorized into

1. **Supervised machine learning –** We need to train our model on a labelled dataset that means we have both raw input data as well as its results. We split our data into a training dataset and test dataset where the training dataset is used to train our model whereas the test dataset acts as new data for predicting results or to see the accuracy of our model. Examples of supervised learning methods are Linear regression, logistic regression, random forest, support vector machine, etc.
2. **Unsupervised machine learning -** Algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data. Examples of Unsupervised learning methods are K-Means clustering, Hierarchical clustering, Principal component analysis, etc.

Let's walkthrough the most important regression algorithms in Supervised machine learning techniques.

## What is Regression

Regression analysis is form of predictive modelling techniques which investigates relationship between one dependent variable (say 'y') and one or more independent variable (say 'X').

Majority of the machine learning algorithms fall under the supervised learning category. It is the process where an algorithm is used to predict a result based on the previously entered values and the results generated from them.

Suppose we have an input variable 'X' and an output variable 'y' where y is a function of X (y=f{X}). Supervised learning reads the value of entered variable 'X' and the resulting variable 'y' so that it can use those results to later predict a highly accurate output data of 'y' from the entered value of 'X'. A regression problem is when the resulting variable contains a real or a continuous value. It tries to draw the line of best fit from the data gathered from a number of points.


## Types of Regression

A. Linear Regression
B. Logistic Regression

## A) LINEAR REGRESSION

Linear Regression is type of regression analysis of independent variables and where there is linear relationship between one or more independent variable (X) and dependent variable (y).

Let's say we have a dataset which contains information about the relationship between '**number of hours studied**' and '**marks obtained**'. A number of students have been observed and their hours of study along with their grades are recorded. This will be our training data. Our goal is to design a model that can predict the marks if number of hours studied is provided. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used to apply for a new data. That is, if we give the number of hours studied by a student as an input, our model should be able to predict their mark with minimum error.

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + .... + \theta_n x_n$$

Where, Y is the predicted value, $\theta_o$ is the bias term or coefficient and $\theta_1$,.....,$\theta n$ are the model parameters, $X_1$, $X_2$,…,Xn are the feature variables/ independent variables.

The above hypothesis can also be represented by

$$Y = \theta^T x$$

where, θ is the model's parameter vector including the bias term $\theta_o$; X is the feature vector with $X_o$ =1

**Y (pred) = $b_0$ + $b_1$*X**

The values $b_0$ and $b_1$ must be chosen so that the error is minimum. If sum of squared error is taken as a metric to evaluate the model, then the goal is to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

If we don't square the error, then the positive and negative points will cancel each other out.

For a model with one predictor,

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

**Exploring $b_1$**

If $b_1 > 0$, then X (predictor) and y (target) have a positive relationship. That is an increase in X will increase y.

If $b_1 < 0$, then X (predictor) and y (target) have a negative relationship. That is an increase in X will decrease y.

**Exploring $b_0$**

If the model does not include X=0, then the prediction will become meaningless with only $b_0$. For example, we have a dataset that relates height(X) and weight(y). Taking x=0 (that is height as 0), will make the equation have only $b_0$ value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.

If the model includes value 0, then '$b_0$' will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

The value of $b_0$ guarantees that the residual will have mean zero. If there is no '$b_0$' term, then the regression will be forced to pass over the origin. Both the regression coefficient and prediction will be biased.

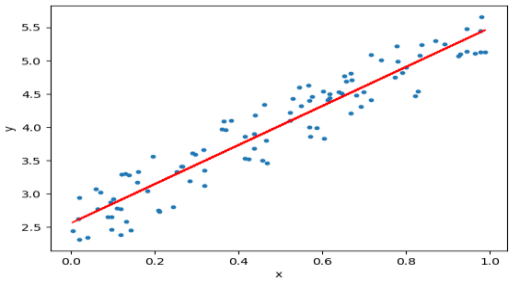
### Working of Linear Regression

Let's look at a scenario where linear regression might be useful: Losing weight.

Let us consider that there's a connection between how many calories you take in and how much you weigh; regression analysis can help you understand that connection. Regression analysis will provide

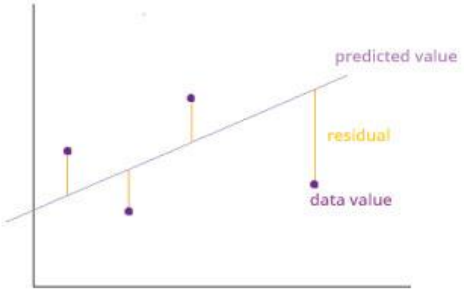
you with a relation which can be visualized into a graph in order to make predictions about your data. For example, if you've been putting on weight over the last few years, it can predict how much you'll weigh in the next ten years if you continue to consume the same amount of calories and burn them at the same rate.

The goal of regression analysis is to create a trend line based on the data you have gathered. This then allows you to determine whether other factors apart from the amount of calories consumed affect your weight, such as the number of hours you sleep, work pressure, level of stress, type of exercises you do etc. Before taking into account, we need to look at these factors and attributes and determine whether there is a correlation between them. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long time duration, extensive data can be collected and the result can be evaluated more accurately.




## How to determine the best fit line?

The best fit line is considered to be the line for which the error between the predicted values and the actual values is minimum. It is also called the *regression line* and the errors are also known as *residuals*. The figure shown below shows the residuals. It can be visualized by the vertical lines from the actual data value to the regression line.




The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

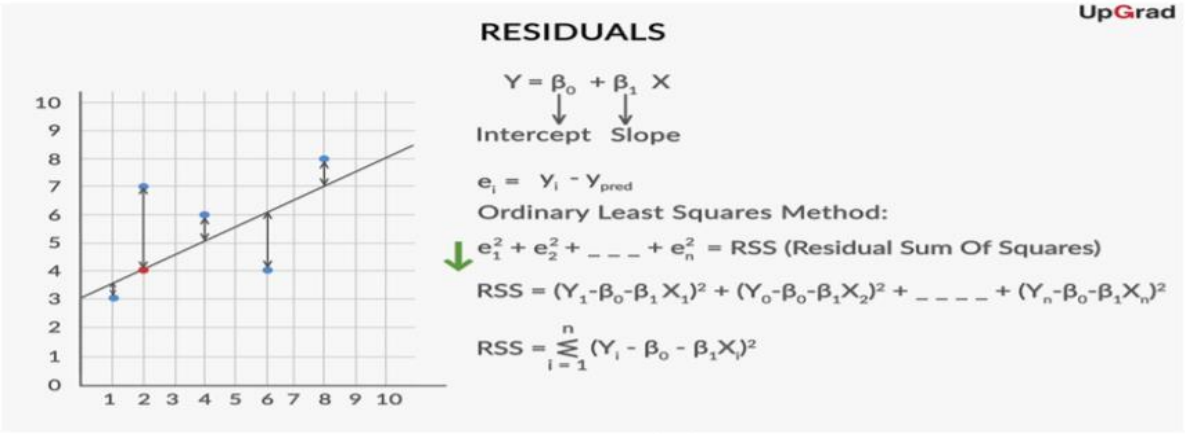



*Figure 4 – Residuals*

## Assumptions in linear regression

If you are planning to use linear regression for your problem then there are some assumptions you need to consider:

- The relation between the dependent and independent variables should be linear.
- The data is homoscedastic, meaning the variance between the results should not be too much.
- The results obtained from an observation should not be influenced by the results obtained from the previous observation.
- The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

**Here are a few features a regression line has:**

- Regression passes through the mean of independent variable (X) as well as mean of the dependent variable (y).
- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)".
- $B_1$ explains the change in y with a change in X by one unit. In other words, if we increase the value of 'X' it will result in a change in value of y.

**Finding a Linear Regression line**

Let's say we want to predict 'y' from 'X' given in the following table and assume they are correlated as $y=B0+B1*X$

| X | y | Predicted y |
|---|---|---|
| 1 | 2 | $B_0+B_1*1$ |
| 2 | 1 | $B_0+B_1*2$ |
| 3 | 3 | $B_0+B_1*3$ |
| 4 | 6 | $B_0+B_1*4$ |
| 5 | 9 | $B_0+B_1*5$ |
| 6 | 11 | $B_0+B_1*6$ |
| 7 | 13 | $B_0+B_1*7$ |
| 8 | 15 | $B_0+B_1*8$ |
| 9 | 17 | $B_0+B_1*9$ |
| 10 | 20 | $B_0+B_1*10$ |
| **Std. Dev. of X** | | 3.02765 |
| **Std. Dev. of y** | | 6.617317 |
| **Mean of X** | | 5.5 |
| **Mean of y** | | 9.7 |
| **Correlation between X & y** | | 0.989938 |

If the Residual Sum of Square (RSS) is differentiated with respect to B0 & B1 and the results equals to zero, we get the following equation:

$B_1$ = Correlation * (Std. Dev. of y/ Std. Dev. of x)

$B_0$ = Mean(Y) – $B_1$ * Mean(X)

Putting values from table 1 into the above equations,

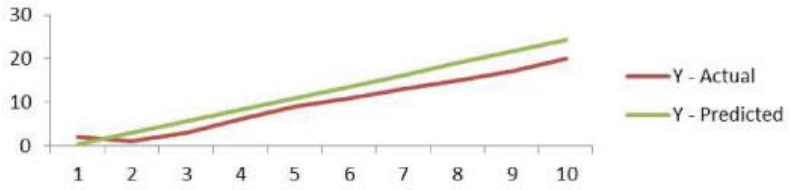
$B_1 = 2.64$

$B_0 = -2.2$

Hence, the least regression equation will become – **Y = -2.2 + 2.64*x**

| X | Y-Actual | Y-Predicted |
|---|---|---|
| 1 | 2 | 0.44 |
| 2 | 1 | 3.08 |
| 3 | 3 | 5.72 |
| 4 | 6 | 8.36 |
| 5 | 9 | 11 |
| 6 | 11 | 13.64 |
| 7 | 13 | 16.28 |
| 8 | 15 | 18.92 |
| 9 | 17 | 21.56 |
| 10 | 20 | 24.2 |

As there are only 10 data points, the results are not too accurate but if we see the correlation between the predicted and actual line, it has turned out to be very high; both the lines are moving almost together and here is the graph for visualizing our predicted values:




## Model Performance

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions. The amount that we consider "not much" entirely depends on the task we want to perform and to what percentage the variation in data can be handled. Here are a few metric tools we can use to calculate error in the model-

**R – Square ($R^2$):** $R^2$ is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as: R 2 = 1 - (RSS / TSS)
Where RSS = Residual Sum of Square and TSS = Total sum of squares

$$R^2 = \frac{TSS - RSS}{TSS}$$

**Total Sum of Squares (TSS):** TSS is a quantity that appears as part of a standard way of presenting results of such an analysis. Sum of squares is a measure of how a data set varies around a central number (like the mean). The Total Sum of Squares tells how much variation there is in the dependent variable.

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

**Residual Sum of Squares (RSS):** In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

(TSS – RSS) measures the amount of variability in the response that is explained by performing the regression.

### Properties of $R^2$

- $R^2$ always ranges between 0 to 1.
- $R^2$ of 0 means that there is no correlation between the dependent and the independent variable.
- $R^2$ of 1 means the dependent variable can be predicted from the independent variable without any error.
- An $R^2$ between 0 and 1 indicates the extent to which the dependent variable is predictable.
- An $R^2$ of 0.20 means that there is 20% of the variance in Y is predictable from X; an $R^2$ of 0.40 means that 40% is predictable; and so on.

### Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). The formula for calculating RMSE is:

$$R^2 = \{(1/N) * \sum [ (x_i - \text{mean}(x)) * (y_j - \text{mean}(y)) ] / (\sigma_x * \sigma_y) \}^2$$

Where N : Total number of observations

When standardized observations are used as RMSE inputs, there is a direct relationship with the correlation coefficient. For example, if the correlation coefficient is 1, the RMSE will be 0, because all of the points lie on the regression line (and therefore there are no errors).

### Mean Absolute Percentage Error (MAPE)

There are certain limitations to the use of RMSE, so analysts prefer MAPE over RMSE which gives error in terms of percentages so that different models can be considered for the task and see how they perform. Formula for calculating MAPE can be written as:

$$\text{RMSE} = \sqrt{\frac{\sum (Y_{Actual} - Y_{Predicted})^2}{N}}$$

Where N : Total number of observations

### Cost Function

Cost function helps to figure out the best possible plots which can be used to draw the line of best fit for the data points. As we want to reduce the error of the resulting value, we change the process of finding out the actual result to a process which can reduce the error between the predicted value and the actual value.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Here, J is the cost function.

The above function is made in this format to calculate the error difference between the predicted values and the plotted/actual values. We take the square of the summation of all the data points and divide it by the total number of actual data points. This cost function J is also called the Mean Squared Error (MSE) function. Using this MSE function we are going to predict values such that the MSE value settles at the minima, reducing the cost function.

## Gradient Descent

Gradient Descent is an optimization algorithm that helps machine learning models to find out paths to a minimum value using repeated steps. Gradient descent is used to minimize a function so that it gives the lowest output of that function. This function is called the Loss Function. The loss function shows us how much error is produced by the machine learning model compared to actual results. Our aim should be to lower the cost function as much as possible. One way of achieving a low cost function is by the process of gradient descent. Complexity of some equations makes it difficult to use, partial derivative of the cost function with respect to the considered parameter can provide optimal coefficient value.

There are mainly 2 types of Linear Regression, namely Simple linear regression and Multiple linear regression.

## SIMPLE LINEAR REGRESSION

The most elementary type of regression model is the simple linear regression which explains the relationship between a one dependent variable and one independent variable using a straight line.

**Residual Analysis:** Simple linear regression models the relationship between the magnitude of one variable and that of a second—for example, as X increases, y also increases. Or as X increases, y decreases. Correlation is another way to measure how two variables are related. The models done by simple linear regression estimate or try to predict the actual result but most often they deviate from the actual result. Residual analysis is used to calculate by how much the estimated value has deviated from the actual result.

**Null Hypothesis and p-value:** During feature selection, null hypothesis is used to find which attributes will not affect the result of the model. Hypothesis tests are used to test the validity of a claim that is made about a particular attribute of the model. This claim that's on trial, in essence, is called the null hypothesis. A p-value helps to determine the significance of the results. p-value is a number between 0 and 1 and is interpreted in the following way:

- A small p-value (less than 0.05) indicates a strong evidence against the null hypothesis, so the null hypothesis is to be rejected.
- A large p-value (greater than 0.05) indicates weak evidence against the null hypothesis, so the null hypothesis is to be failed to reject Null hypothesis.
- A p-value very close to the cut-off (equal to 0.05) is considered to be marginal.

## Ordinary Least Square

Ordinary Least Squares (OLS), also known as Ordinary least squares regression or least squared errors regression is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters for a linear function, the goal of which is to minimize the sum of the squares of the difference of the observed variables and the dependent variables i.e. it tries to attain a relationship between them.

There are two types of relationships that may occur: linear and curvilinear. A linear relationship is a straight line that is drawn through the central tendency of the points; whereas a curvilinear relationship is a curved line. Association between the variables are depicted by using a scatter plot. The relationship could be positive or negative, and result variation also differs in strength.

The advantage of using Ordinary Least Squares regression is that it can be easily interpreted and is highly compatible with recent computers' built-in algorithms from linear algebra. It can be used to apply to problems with lots of independent variables which can efficiently conveyed to thousands of data points. In Linear Regression, OLS is used to estimate the unknown parameters by creating a model which will minimize the sum of the squared errors between the observed data and the predicted one.

## Regularization

Regularization is a type of regression that is used to decrease the coefficient estimates down to zero. This helps to eliminate the data points that don't actually represent the true properties of the model, but have appeared by random chance. The process is done by identifying the points which have deviated from the line of best-fit by a large extent. Earlier we saw that to estimate the regression

coefficients β in the least squares method, we must minimize the term Residual Sum of Squares (RSS). Let the RSS equation in this case be:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 x_i + \beta_0)^2$$

The general linear regression model can be expressed using a condensed formula:

$$Y = X * \beta$$

Here, $\beta=[\beta_0, \beta_1, \ldots. \beta_p]$

The RSS value will adjust the coefficient, β based on the training data. If the resulting data deviates too much from the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

**MULTIPLE LINEAR REGRESSION**

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Consider example of sales prediction using TV marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e.TV marketing, radio marketing, and newspaper marketing. We need to consider multiple variables as just one variable alone might not be good enough to explain the feature variable, in this case, Sales. The table below shows how adding a variable helped increase the R-squared that we had obtained by using just the TV variable.

| TV | Radio | Newspaper | Sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |

| Predictors | R squared |
|---|---|
| TV | 0.816 |
| Radio | 0.112 |
| Newspaper | 0.058 |

| | |
|---|---|
| TV + Newspaper | 0.836 |
| TV + Radio | 0.910 |

So we see that adding more variables increases the R-squared and it might be a good idea to use multiple variables to explain a feature variable. Basically:

1. Adding variables helped add information about the variance in Y.
2. In general, we expect explanatory power to increase with increase in variables.

Hence, this brings us to multiple linear regression which is just an extension to simple linear regression. The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from Simple Linear Regression still hold
    - Zero mean, independent, Normally distributed error terms that have constant variance
    - The inference part in multiple linear regression also, largely, remains the same

**Moving from SLR to MLR**

In simple linear regression we use a single independent variable to predict the value of a dependent variable whereas in multiple linear regression two or more independent variables are used to predict

the value of a dependent variable. The difference between the two is the number of independent variables. In both cases there is only a single dependent variable.
Although, most of the ideas in simple and multiple linear regression are the same, there are a few new considerations that you need to make when moving to multiple linear regression, such as:
a. Adding more isn't always helpful, model may 'overfit' by becoming too complex
b. Multicollinearity - Associations between predictor variables

**Overfitting:** When you add more and more variables, for example, let's say you keep on increasing the degree of the polynomial function fitting the data, your model might end up memorizing all the data points in the training set. This will cause major problems with generalisation, i.e. now when the model runs on the test data, the accuracy will drop tremendously since, it doesn't generalise well. This is a classical symptom of overfitting.

**Multicollinearity:** Multicollinearity is the effect of having related predictors in the multiple linear regression model. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, i.e. some of these variables might completely explain some other independent variable in the model due to which the presence of that variable in the model is redundant. So in order to know, where the effect on the feature variable is coming from, we need to drop some of these related independent variables.

Multicollinearity tells us the strength of the relationship between independent variables. Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable. VIF (Variance Inflation Factor) is used to identify the Multicollinearity. If VIF value is greater than 5, we need to exclude that variable from our model.

There are certain reasons why multicollinearity occurs:

- It is caused by an inaccurate use of dummy variables.
- It is caused by the inclusion of a variable which is computed from other variables in the data set.
- Multicollinearity can also result from the repetition of the same kind of variable.
- Generally occurs when the variables are highly correlated to each other.

There are two ways to detect multicollinearity in a model:

● **Correlations:** Looking at pairwise correlations between the independent variables can sometimes be useful to detect multicollinearity.

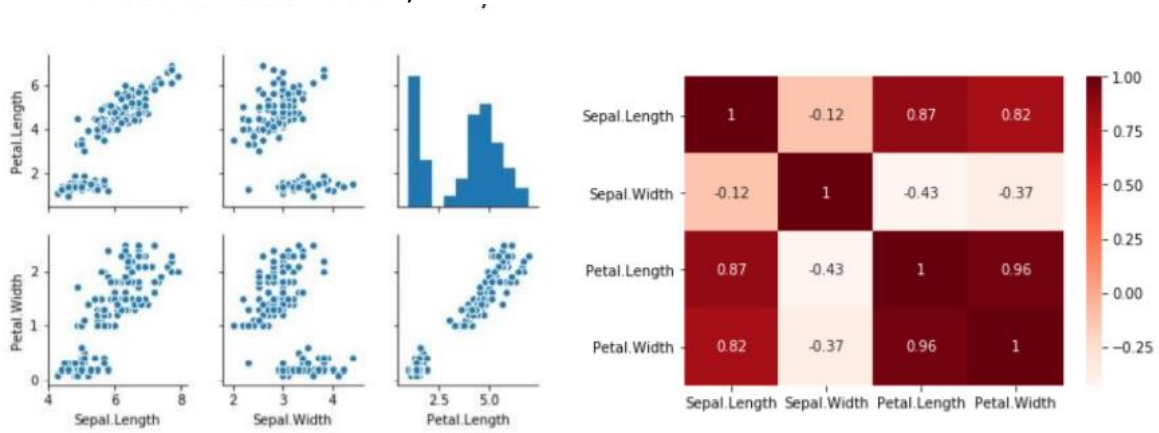



*Fig 13 - Pairwise Correlations*

From the images above, plotted for Iris dataset, you can clearly see that some of the pair of variables, such as, petal width and septal length, petal width and petal length, etc. are highly correlated. Hence, when the model is built, one of the variables from each of these pairs of variables might turn out to be redundant for the model.

● **Variance Inflation Factor (VIF):** Now, looking at correlations might not always be useful as it is possible that just one variable might not completely explain some other variable but some of the

variables combined might be able to do that. To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. We always consider VIF is always < 5.
Now, after any multicollinearity has been detected in the model, we need to deal with it appropriately in order to avoid building an unnecessarily complex model with a lot of redundant variables. The few methods to deal with multicollinearity are:

### Dropping Variables

Drop the variable that is highly correlated with others and that we need to check with combination of p-value and VIF value. We need to drop the independent variable which has p-value is greater than 0.05 and VIF value is greater than 5 until we get a balanced model. Also, most important is while dropping the independent variables, we shouldn't drop all the variables which has p-value > 0.5, follow one by one and continuously check the p-value after dropping the variable until we get good predictive model.

**Feature Scaling:** Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons: **(i).** Ease of interpretation **(ii).** Faster convergence for gradient descent methods.
You can scale the features using two very popular method:
**(i) Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

**(ii) MinMax Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc

### Handling Categorical Variable

In simple linear regression, we worked with just numeric variables. But when you have multiple variables, there might be some categorical variables that might turn out to be useful for the model. So it is essential to handle these variables appropriately in order to get a good model. One way to deal with them is creating dummy variables. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. See the below example to get a clearer idea.

| Value | Indicator Variable |
|---|---|
| Gender | Female |
| Male | 0 |
| Female | 1 |

| Value | Indicator Variable | |
|---|---|---|
| Furnishing Status | furnished | semi-furnished |
| furnished | 1 | 0 |
| semi-furnished | 0 | 1 |
| unfurnished | 0 | 0 |

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

*Fig 14 - Dummy variables*

Since, a multiple linear regression can be built with different combinations of the variables present, model comparison and hence, selection of the best model becomes extremely essential. The key aspect while selecting the best model is the trade-off between selecting the model explaining the variance best and the model which is fairly simple. So to implement this idea, we need a few parameters apart from the original ones (like R-squared) that would test the goodness of the model as well as penalise the model for using more number of predictor variables. Hence, two new parameters come into picture to assess a multiple linear regression model:

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

$$AIC = n * log(\frac{RSS}{n}) + 2p$$

These parameters are useful for selecting the best model that is fairly simple as well as explains a decent amount of variance. Apart from these, there is another parameter called BIC, which is quite similar to AIC, the only difference being that it penalises the model more for adding more variables.

**FEATURE SELECTION**

Feature selection is the automatic selection of attributes for our data that are most relevant to the predictive model we are working on. It seeks to reduce the number of attributes in the dataset by eliminating the features which are not required for the model construction. Feature selection does not totally eliminate an attribute which is considered for the model, rather it mutes that particular characteristic and works with the features which affects the model.

Feature selection method aids your mission to create an accurate predictive model. It helps you by choosing features that will give you as good or better accuracy whilst requiring less data. Feature selection methods can be used to identify and remove unnecessary, irrelevant and redundant attributes from the data that do not contribute to the accuracy of the model or may even decrease the accuracy of the model. Having fewer attributes is desirable because it reduces the complexity of the model, and a simpler model is easier to understand, explain and to work with.

So let's talk about the various methods for optimal feature selection:

1. Try all possible combinations (2 p models for p features) → Time consuming and practically unfeasible
2. Manual Feature Elimination → Build model, Drop features that are least helpful in prediction (high p-value), Drop features that are redundant (using correlations, VIF), Rebuild model and repeat these steps.
3. Automated Approach → Recursive Feature Elimination(RFE), Forward/Backward/Stepwise Selection based on AIC.

It is generally recommended that you follow a balanced approach, i.e., use a combination of automated (coarse tuning) + manual (fine tuning) selection in order to get an optimal model.
This is all about linear regression.

## B. LOGISTIC REGRESSION

Logistic regression is a supervised learning classification algorithm (binary classification algorithm) used to predict the probability of a target variable/dependent variable. There would be only 2 possible values of dependent variable yes or no, 0 or 1, success or failure. It is one of the most frequently used machine learning algorithms for binary classifications that translates the input to 0 or 1. For example, 0- negative class and 1- positive class.

Some examples of classification are mentioned below:

- A bank wants to predict, based on some variables, whether a particular customer will default on a loan or not.
- A factory manager wants to predict, based on some variables, whether a particular machine will break down in the next month or not.
- Google's backend wants to predict, based on some variables, whether an incoming email is spam or not

The reasons why linear regressions cannot be used in case of binary classification are as follows:

- **Distribution of error terms:** The distribution of data in the case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In the case of binary classification, this assumption does not hold true.
- **Model output:** In linear regression, the output is continuous. In the case of binary classification, an output of a continuous value does not make sense. For binary classification problems, linear regression may predict values that can go beyond 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different classes, then its range should be restricted to 0 and 1. As the logistic regression model can output probabilities with logistic/sigmoid function, it is preferred over linear regression.
- **Variance of Residual errors:** Linear regression assumes that the variance of random errors is constant. This assumption is also violated in the case of logistic regression.

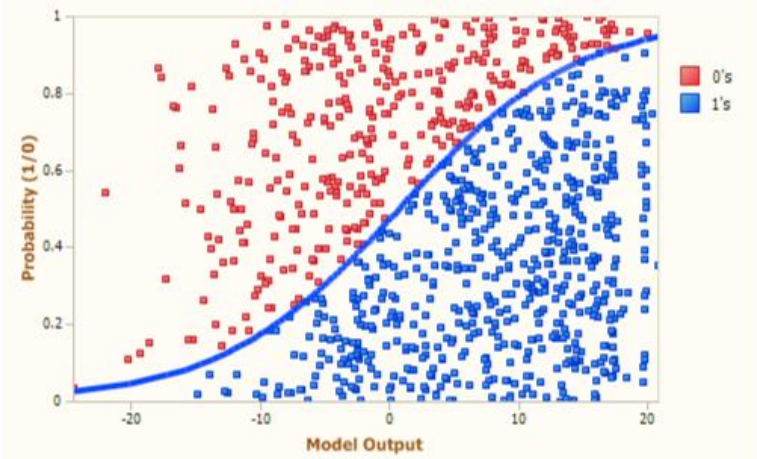
### What is Logistic Regression

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable has a binary solution. Similar to all other types of regression systems, Logistic Regression is also a type of predictive regression system. Logistic regression is used to evaluate the relationship between one dependent binary variable and one or more independent variables. It gives discrete outputs ranging between 0 and 1.

A simple example of Logistic Regression is: Does calorie intake, weather, and age have any influence on the risk of having a heart attack? The question can have a discrete answer, either "yes" or "no".

### When to use Logistic Regression?

Logistic Regression is used when the input needs to be separated into "two regions" by a linear boundary. The data points are separated using a linear line as shown:



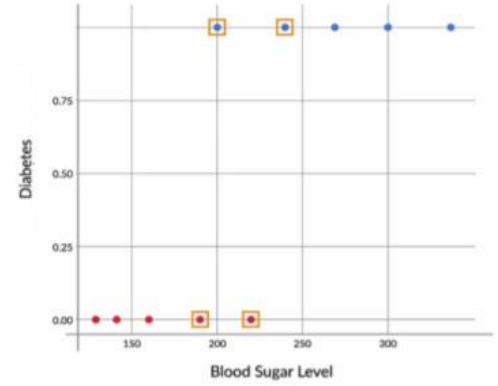
Based on the number of categories, Logistic regression can be classified as:

a) **Binomial:** target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
b) **Multinomial:** target variable can have 3 or more possible types which are not ordered(i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.
c) **Ordinal:** it deals with target variables with ordered categories. For example, a test score can be categorized as: “very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.

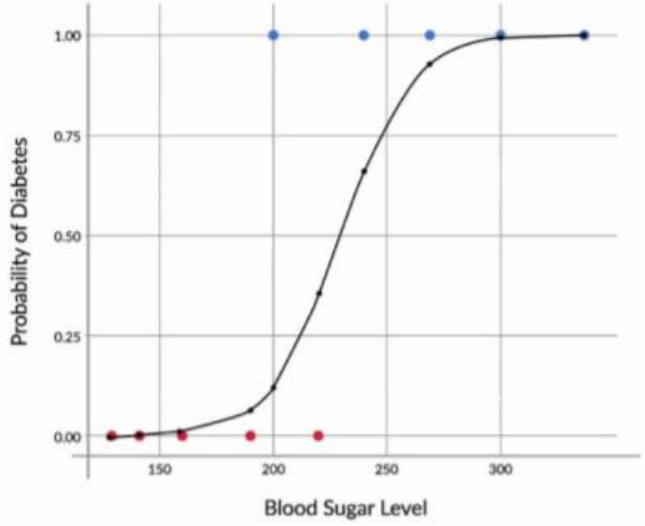
**Example**

In this example, we try to predict whether a person has diabetes or not, based on that person’s blood sugar level. We knew why a simple boundary decision approach does not work very well for this example. It would be too risky to decide the class blatantly on the basis of cut off, as especially in the middle, the patients could basically belong to any class, diabetic or non-diabetic.

| Blood Sugar Level | 190 | 240 | 300 | 160 | 200 | 269 | 129 | 141 | 220 | 337 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diabetes | No | Yes | Yes | No | Yes | Yes | No | No | No | Yes |




Hence, you learnt it is better, actually to talk in terms of probability. One such curve which can model the probability of diabetes very well, is the sigmoid curve.



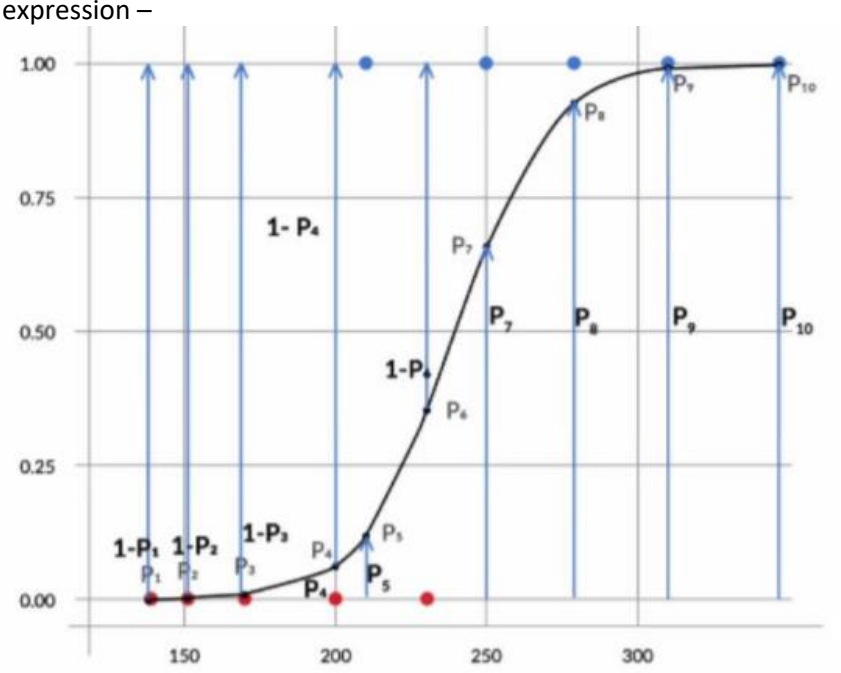

Its equation is given by the following expression –

$$P(\text{Diabetes}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

**Likelihood**

The next step, just like linear regression, would be to find the best fit curve. Hence, we knew that in order to find the best fit sigmoid curve, we need to vary β0 and β1 until we get the combination of

beta values that maximises the likelihood. For the diabetes example, likelihood is given by the expression –




**Likelihood = (1-P1)(1-P2)(1-P3)(1-P4)(P5)(1-P6)(P7)(P8)(P9)(P10)**

Generally, Likelihood function is the product of –

[(1-Pi)(1-Pi) ------ for all non-diabetics --------] * [(Pi)(Pi) -------- for all diabetics -------]

This process, where we vary the betas, until we find the best fit curve for probability of diabetes, is called logistic regression.

## ODDS and Log ODDS

This best fit line function gives the relationship between P (the probability of target variable) and X (the independent variables). But, this form of the function is not very intuitive i.e the relationship between P and X is so complex that it becomes difficult to understand what kind of trend exists between the two. However, on converting the function to a slightly different form, one can achieve a much more intuitive relationship. Upon taking the natural logarithm on both sides of the equation one gets,

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x \quad where, \ \frac{P}{1-P} = Odds \quad and \quad \log\left(\frac{P}{1-P}\right) = Log\ Odds$$

Odds are defined as the ratio of probability of success to the probability of failure. So, if the odds of success is 4 (0.8/0.2) , it shows that the odds of success ( 80% ) has an accompanying odds of failure ( 20% ). Whereas , Log odds is the logarithm of the odds i.e. Log(4) = 1.386

Multivariate Logistic Regression

Multivariate Logistic regression is just an extension of the Univariate Logistic regression. So, the form is identical to Univariate Logistic regression, but with more than one covariate. The logit equation is given by

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots)}}$$

## Assumptions on Logistic Regression

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.

- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome.
- There is no influential values (extreme values or outliers) in the continuous predictors.
- There is no high intercorrelations (i.e. multicollinearity) among the predictors.

To improve the accuracy of the model, we should make sure that these assumptions hold true for our dataset.

### How is OLS different from MLE?

Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

Ordinary Least Squares (OLS) also called the linear least squares is a method to approximately determine the unknown parameters of a linear regression model. Ordinary least squares is obtained by minimizing the total squared vertical distances between the observed responses within the dataset and the responses predicted by the linear approximation (represented by the line of best fit or regression line). The resulting estimator can be represented using a simple formula.

For example, let's say you have a set of equations which consist of several equations with unknown parameters. The ordinary least squares method may be used because this is the most standard approach in finding the approximate solution to your overly determined systems. In other words, it is your overall solution in minimizing the sum of the squares of errors in your equation. Data that best fits the ordinary least squares minimizes the sum of squared residuals. Residual is the difference between an observed value and the predicted value provided by a model.

Maximum likelihood estimation, or MLE, is a method used in estimating the parameters of a statistical model, and for fitting a statistical model to data. If we want to find the height measurement of every basketball player in a specific location, maximum likelihood estimation can be used. If we could not afford to measure all of the basketball players' heights, the maximum likelihood estimation can come in very handy. Using the maximum likelihood estimation, you can estimate the mean and variance of the height of your subjects. The MLE would set the mean and variance as parameters in determining the specific parametric values in a given model.

To sum it up, the maximum likelihood estimation covers a set of parameters which can be used for predicting the data needed in a normal distribution. A given, fixed set of data and its probability model would likely produce the predicted data. The MLE would give us a unified approach when it comes to the estimation. But in some cases, we cannot use the maximum likelihood estimation because of recognized errors or the problem actually doesn't even exist in reality.

### Strength of Logistic Regression Model

The strength of any logistic regression model can be assessed using various metrics. These metrics usually provide a measure of how well the observed outcomes are being replicated by the model, based on the proportion of total variation of outcomes explained by the model. The various metrics used to measure a model are,

1. Confusion Matrix
2. ROC Curve
3. Gain and Lift Chart
4. KS - Statistic
5. Gini Coefficient and will discuss 2 mostly use common measures.

### Confusion Matrix

A confusion matrix is an n × n matrix, where n is the number of classes being predicted. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. Confusion Matrix gives insight not only into the errors being made by your classifier but more importantly the types of errors that are being made. It is this breakdown that overcomes the limitation of using classification accuracy alone.

The following figure represents a confusion matrix details.

| Confusion Matrix | | Target | | *Precision* *TP/(TP+FP)* |
|---|---|---|---|---|
| | | **Positive** | **Negative** | |
| **Model** | **Positive** | True Positive | False Positive | *Positive Predictive Value* *TP/(TP+FP)* |
| | **Negative** | False Negative | True Negative | *Negative Predictive Value* *TN/(FN+TN)* |
| *Recall* *TP/(TP+FN)* | | *Sensitivity* *TP/(TP+FN)* | *Specificity* *TN/(FP+TN)* | **Accuracy** **(TP+TN)/(TP+FP+FN+TN)** |

**a) Accuracy** of the model is the proportion of the total number of predictions that were correct.
Accuracy = (TP + TN) / (TP + FP + FN + TN)
For a model with an accuracy of about 90% (which looks good), on revaluation of the confusion matrix, one could see that there were still a lot of misclassifications present. Thus, other new discriminative metrics were brought in.
**b) Positive Predictive Value or Precision:** Proportion of positive cases correctly identified.
Precision is used as a measure to calculate the success of predicted values to the values which were supposed to be successful. Precision is used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system. It shows how many of the positively classified were relevant. A test can cheat and maximize this by only returning positive on one result it's most confident in.
Positive Predictive Value or Precision = TP/(TP + FP)
**c) Negative Predictive Value:** Proportion of negative cases correctly identified.
Negative Predictive Value = TN/(TN + FN)
**d) Sensitivity / Recall:** Proportion of actual positive cases correctly identified.
It measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate. It shows how good a test is at detecting the positives. A test can cheat and maximize this by always returning "positive".
Sensitivity or Recall = TP/(TP + FN)
**e) Specificity:** Proportion of actual negative cases correctly identified.
It (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. It shows how good a test is at avoiding false alarms. A test can cheat and maximize this by always returning "negative".
Specificity = TN/(TN + FP)
**f) True Positive Rate:** Proportion of actual positive cases correctly identified.
True Positive Rate = TP/(TP + FN)
**g) False Positive Rate:** Proportion of actual negative cases incorrectly identified.
Specificity = FP/(FP + TN)
**h) F1 Score:** It is the harmonic mean of Precision and Recall.
F1 Score = 2 × (Precision × Recall)/(Precision + Recall)

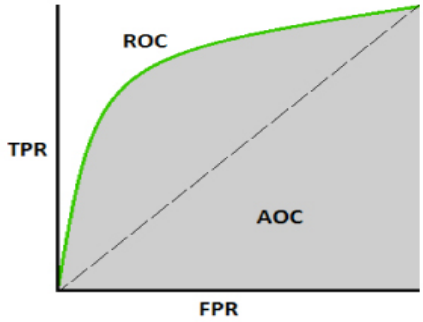
## ROC Curve

The ROC or Receiver Operating Characteristics curve is the plot between True Positive Rate and the False Positive Rate, or simply, a trade off between sensitivity and specificity. The biggest advantage of using the ROC curve is that it is independent of the change in proportion of responders. This is because it has both the axes derived out from the columnar calculations of confusion matrix, where the numerator and denominator of both x and y axis change on a similar scale for any change in the

response rate. The ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{TP}{TP+FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positives}{True\ Negatives + False\ Positives} = \frac{FP}{TN+FP}$$




There are a number of methods of evaluating whether a logistic model is a good model. One such way is sensitivity and specificity. Sensitivity and Specificity are statistical measures of the performance of a binary classification test.The above figure shows the more the ROC curve is towards the upper-left corner (i.e the more is the area under the curve), the better is the model.
As sometimes the accuracy, sensitivity and specificity trade off, that when the probability thresholds are very low, the sensitivity is very high and the specificity is very low. Similarly, for larger probability thresholds, the sensitivity values are very low but the specificity values are very high. One could choose any cut-off point based on which of these metrics is required to be high (like if one wants to capture the positives better than some accuracy could be let off for the sake of higher sensitivity and a lower cut-off be chosen). It is completely dependent on the situation. But the optimal cut-off point (where accuracy, sensitivity and specificity meet) can give a fair idea of how the thresholds should be chosen.

## Pros and Cons of Logistic Regression

Many of the pros and cons of the linear regression model also apply to the logistic regression model. Although Logistic regression is used widely by many people for solving various types of problems, it fails to hold up its performance due to its various limitations and also other predictive models provide better predictive results.

**Pros**

- The logistic regression model not only acts as a classification model, but also gives you probabilities. This is a big advantage over other models where they can only provide the final classification. Knowing that an instance has a 99% probability for a class compared to 51% makes a big difference. Logistic Regression performs well when the dataset is linearly separable.
- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative). We see that Logistic regression is easier to implement, interpret and very efficient to train.

**Cons**

- Logistic regression can suffer from complete separation. If there is a feature that would perfectly separate the two classes, the logistic regression model can no longer be trained. This is because the weight for that feature would not converge, because the optimal weight would be infinite. This is really a bit unfortunate, because such a feature is really very useful. But you do not need machine learning if you have a simple rule that separates both classes. The

problem of complete separation can be solved by introducing penalization of the weights or defining a prior probability distribution of weights.

- Logistic regression is less prone to overfitting but it can overfit in high dimensional datasets and in that case, regularization techniques should be considered to avoid over-fitting in such scenarios.

| | **Linear Regression** | **Logistic Regression** |
|---|---|---|
| **Outcome** | In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values. | In logistic regression, the outcome (dependent variable) has only a limited number of possible values. |
| **The dependent variable** | Linear regression is used when your response variable is continuous. For instance, weight, height, number of hours, etc. | Logistic regression is used when the response variable is categorical in nature. For instance, yes/no, true/false, red/green/blue, 1st/2nd/3rd/4th, etc. |
| **The independent variable** | In Linear Regression, the independent variables can be correlated with each other. | In logistic Regression, the independent variables should not be correlated with each other. (no multi-collinearity) |
| **Equation** | Linear regression gives an equation which is of the form Y = mX + C, means equation with degree 1. | Logistic regression gives an equation which is of the form Y = eX + e-X. |
| **Coefficient interpretation** | In linear regression, the coefficient interpretation of independent variables are quite straightforward (i.e. holding all other variables constant, with a unit increase in this variable, the dependent variable is expected to increase/decrease by xxx). | In logistic regression, depends on the family (binomial, Poisson, etc.) and link (log, logit, inverse-log, etc.) you use, the interpretation is different. |
| **Error minimization technique** | Linear regression uses ordinary least squares method to minimise the errors and arrive at a best possible fit, while logistic regression uses maximum likelihood method to arrive at the solution. | Logistic regression is just the opposite. Using the logistic loss function causes large errors to be penalized to an asymptotic constant. |